The recognition of human movement using temporal templates

Article *in* IEEE Transactions on Pattern Analysis and Machine Intelligence · April 2001

DOI: 10.1109/34.910878 · Source: IEEE Xplore

Aaron Bobick
Washington University in St. Louis
213 PUBLICATIONS **18,186** CITATIONS

The Recognition of Human Movement Using Temporal Templates

Aaron F. Bobick, *Member*, *IEEE Computer Society*, and
James W. Davis, *Member*, *IEEE Computer Society*

Abstract—A new view-based approach to the representation and recognition of human movement is presented. The basis of the representation is a *temporal template*—a static vector-image where the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence. Using aerobics exercises as a test domain, we explore the representational power of a simple, two component version of the templates: The first value is a binary value indicating the presence of motion and the second value is a function of the recency of motion in a sequence. We then develop a recognition method matching temporal templates against stored instances of views of known actions. The method automatically performs temporal segmentation, is invariant to linear changes in speed, and runs in real-time on standard platforms.

Index Terms—Motion recognition, computer vision.

1 Introduction

There is a rich tradition in computer vision of studying image sequences, an early survey can be found in [1]. But recently, the focus of research is less on the measurement of image or camera motion and more on the labeling of the action taking place in the scene. This shift has been triggered not only by the availability of the computational resources, but also the interest in applications, such as wireless interfaces (e.g., [13]) and interactive environments [21], [5]. The fundamental question is no longer "How are things (pixels or cameras) moving?" but, rather "What is happening?"

Unfortunately, this new labeling problem is not as well-defined as the previously addressed questions of geometry. Bobick [6] considers the range of motion interpretation problems and proposes a taxonomy of approaches. At the top and intermediate levels—action and activity, respectively—are situations in which knowledge other than the immediate motion is required to generate the appropriate label. The most primitive level, however, is *movement*—a motion whose execution is consistent and easily characterized by a definite space-time trajectory in some feature space. Such consistency of execution implies that for a given viewing condition there is consistency of appearance. Put simply, movements can be described by their appearance.

This paper presents a novel, appearance-based approach to the recognition of human movement. Our work stands in contrast to many recent efforts to recover the full three-dimensional reconstruction of the human form from image sequences, with the presumption that such information would be useful and perhaps even necessary to interpret the motion (e.g., [23]). Instead, we develop a view-based approach to the representation and recognition of movement that is designed to support the direct recognition of the motion itself.

1.1 A Motivating Example

Fig. 1 illustrates the motivation for the work described here and for earlier work, which attempted to exploit similar motion information through a different computational mechanism [7]. Presented are frames of an extremely low resolution sequence in which a subject is performing a normally trivially recognizable movement. Despite the almost total lack of of recognizable features in the static imagery, the movement is easily recognized when the sequence is put in motion on a screen.

This capability of the human vision system argues for recognition of movement directly from the motion itself, as opposed to first reconstructing a three-dimensional model of a person and then recognizing the motion of the model as advocated in [16], [23], [24]. In [7], we first proposed a representation and recognition theory that decomposed motion-based recognition into first describing *where* there is motion (the spatial pattern) and then describing *how* the motion is moving.

In this paper, we continue to develop this approach. We first present the construction of a binary *motion-energy image* (MEI) which represents where motion has occurred in an image sequence. Next, we generate a *motion-history image* (MHI) which is a scalar-valued image where intensity is a function of recency of motion. Taken together, the MEI and MHI can be considered as a two component version of a *temporal template*, a vector-valued image where each component of each pixel is some function of the motion at that pixel location. These view-specific templates are matched against the stored models of views of known movements. To evaluate the power of the representation,

- *A.F. Bobick is with the College of Computing, Georgia Tech, 801 Atlantic Dr., Atlanta, GA 30332. E-mail: afb@cc.gatech.edu.*
- *J.W. Davis is with the Computer and Information Science Department, The Ohio State University, 583 Dreese Labs, 2015 Neil Ave., Columbus, OH 43210. Email: jwdavis@cis.ohio-state.edu.*

Manuscript received 10 Nov. 1998; revised 16 Aug. 2000; accepted 10 Oct. 2000.
Recommended for acceptance by M.J. Black.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 108209.

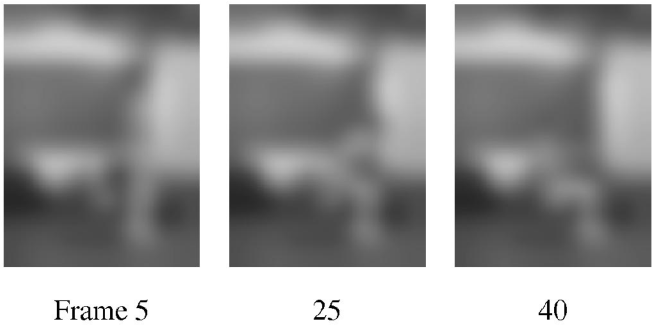


Fig. 1. Selected frames from video of someone performing some movement. Even with almost no structure present in each frame people, can trivially recognize the motion as someone sitting.

we examine the discrimination power on a set of 18 aerobic exercises. Finally, we present a recognition method that automatically performs temporal segmentation, is invariant to linear changes in speed, and runs in real-time on a standard platform.

2 Prior Work

The number of approaches to recognizing motion, particularly human movement, has recently grown at a tremendous rate. Aggarwal and Cai recently provided an extensive survey on the machine analysis of human motion [2]. That paper covers not only issues of recognition, but also of general model recovery (i.e., the three-dimensional structure of the body at each point in time). Because the focus here is on recognition, we will only summarize the model construction techniques, concentrating on recognition strategies that would exploit such information. We divide the prior work into generic model recovery, appearance-based models, and direct motion-based recognition.

2.1 Generic Human Model Recovery

The most common technique for attaining the three-dimensional information of movement is to recover the pose of the person or object at each time instant using a three-dimensional model. The model fitting is driven by attempting to minimize a residual measure between the projected model and object contours (e.g., edges of body in the image). This generally requires a strong segmentation of foreground/background and also of the individual body parts to aid the model alignment process. It is difficult to imagine such techniques could be extended to the blurred sequence of Fig. 1.

For example, Rehg and Kanade [23] used a 27 degree-of-freedom (DOF) model of a human hand in their system called "Digiteyes." Local image-based trackers are employed to align the projected model lines to the finger edges against a solid background. The work of Goncalves et al. [15] promoted three-dimensional tracking of the human arm against a uniform background using a two cone arm model and a single camera. Though it may be possible to extend their approach to the whole body as claimed, it seems unlikely that it is appropriate for nonconstrained human motion with self-occlusion. Hogg [16] and Rohr [24] used a full-body cylindrical model for tracking walking humans in natural scenes. Rohr incorporates a 1 DOF pose parameter to aid in the model fitting. All the poses in a walking action are indexed by a single number. Here, there is only a small subset of poses which can exist. Gavrila and Davis [14] also used a full-body model (22 DOF, tapered superquadrics) for tracking human motion against a complex background. For simplifying the edge detection in cases of self-occlusion, the user is required to wear a tight-fitting body suit with contrasting limb colors.

One advantage of having the recovered model is the ability to estimate and predict the feature locations, for instance, edges, in the following frames. Given the past history of the model configurations, prediction is commonly attained using Kalman filtering [24], [23], [15] and velocity constraints [14].

Because of the self-occlusions that frequently occur in articulated objects, some systems employ multiple cameras and restrict the motion to small regions [23], [14] to help with projective model occlusion constraints. A single camera is used in [16], [15], [24], but the actions tracked in these works had little deviation in the depth of motion. Acquiring the three-dimensional information from image sequences is currently a complicated process, many times necessitating human intervention or contrived imaging environments.

2.1.1 *Three-Dimensional Movement Recognition*

As for action recognition, Campbell and Bobick [8] used a commercially available system to obtain three-dimensional data of human body limb positions. Their system exploits redundancies that exist for particular actions and performs recognition using only the information that varies between actions. This method examines the relevant parts of the body, as opposed to the entire body data. Siskind [26] similarly used known object configurations. The input to his system consisted of line-drawings of a person, table, and ball. The positions, orientations, shapes, and sizes of the objects are known at all times. The approach uses support, contact, and attachment primitives and event logic to determine the actions of dropping, throwing, picking up, and putting down. These two approaches address the problem of recognizing actions when the precise configuration of the person and environment is known while the methods from the previous section concentrate on the recovery of the object pose.

2.2 Appearance-Based Models

In contrast to the three-dimensional reconstruction and recognition approaches, others attempt to use only the two-dimensional appearance of the action (e.g., [3], [10], [9], [29]). View-based representations of two-dimensional statics are used in a multitude of frameworks, where an action is described by a sequence of two-dimensional instances/poses of the object. Many methods require a normalized image of the object (usually with no background) for representation.

For example, Cui et al. [9], Darrell and Pentland [10] and, also, Wilson and Bobick [27] present results using actions (mostly hand gestures), where the actual grayscale images (with no background) are used in the representation for the action. Though hand appearances remain fairly similar over a wide range of people, with the obvious exception of skin color, actions that include the appearance of the total body are not as visually consistent across different people due to obvious natural variations and different clothing. As opposed to using the actual raw gray-scale image, Yamato et al. [29] examines body silhouettes, and Akita [3] employs body contours/edges. Yamato utilizes low-level silhouettes of human actions in a Hidden Markov Model (HMM) framework, where binary silhouettes of background-subtracted images are vector quantized and used as input to the HMMs. In Akita's work [3], the use of edges and some simple two-dimensional body configuration knowledge (e.g., the arm is a protrusion out from the torso) are used to determine the body parts in a hierarchical manner (first, find legs, then head, arms, trunk) based on stability. Individual parts are found by chaining local contour information. These two approaches help alleviate *some* of the variability between people but introduce other problems, such as the disappearance of movement that happens to be within the silhouetted region and also the varying amount of contour/edge information that arises when the background or clothing is high versus low frequency (as in most natural scenes). Also, the problem of examining the entire body, as opposed to only the desired regions, still exists, as it does in much of the three-dimensional work.

Whether using two- or three-dimensional structural information, many of the approaches discussed so far consider an action to be comprised of a sequence of static poses of an object. Underlying all of these techniques is the requirement that there be individual features or properties that can be extracted and tracked from each frame of the image sequence. Hence, motion understanding is really accomplished by recognizing a sequence of static configurations. This understanding generally requires previous recognition and segmentation of the person [22]. We now consider recognition of action within a motion-based framework.

2.3 Motion-Based Recognition

Direct motion recognition [22], [25], [20], [4], [28], [26], [12], [7] approaches attempt to characterize the motion itself without reference to the underlying static poses of the body. Two main approaches include the analysis of the body region as a single "blob-like" entity and the tracking of predefined *regions* (e.g., legs, head, mouth) using motion instead of structural features.

Of the "blob-analysis" approaches, the work of Polana and Nelson [22], Shavit and Jepson [25] and, also, Little and Boyd [20] are most applicable. Polana and Nelson use repetitive motion as a strong cue to recognize cyclic walking motions. They track and recognize people walking in outdoor scenes by gathering a feature vector, over the entire body, of low-level motion characteristics (optical-flow magnitudes) and periodicity measurements. After gathering training samples, recognition is performed using a nearest centroid algorithm. By assuming a fixed height and velocity of each person, they show how their approach may be extended to tracking multiple people in simple cases. Shavit and Jepson also take an approach using the gross overall motion of the person. The body, an animated silhouette figure, is coarsely modeled as an ellipsoid. Optical flow measurements are used to help create a phase portrait for the system, which is then analyzed for the force, rotation, and strain dynamics. Similarly, Little and Boyd recognize people walking by analyzing the motion associated with two ellipsoids fit to the body. One ellipsoid is fit using the motion region silhouette of the person and the other ellipsoid is fit using motion magnitudes as weighting factors. The relative phase of various measures (e.g., centroid movement, weighted centroid movement, torque) for each of the ellipses over time characterizes the gait of several people.

There is a group of work which focuses on motions associated with facial expressions (e.g., characteristic motion of the mouth, eyes, and eyebrows) using region-based motion properties [28], [4], [12]. The goal of this research is to recognize human facial expressions as a dynamic system, where the motion of interest regions (locations known a priori) is relevant. These approaches characterize the expressions using the underlying motion properties rather than represent the action as a sequence of poses or configurations. For Black and Yacoob [4] and, also, Yacoob and Davis [28], optical flow measurements are used to help track predefined polygonal patches placed on interest regions (e.g., mouth). The parameterization and location relative to the face of each patch was given a priori. The temporal trajectories of the motion parameters were qualitatively described according to positive or negative intervals. Then these qualitative labels were used in a rule-based, temporal model for recognition to determine expressions, such as anger or happiness.

Ju et al. [19] have extended this work with faces to include tracking the legs of a person walking. As opposed to the simple, independent patches used for faces, an articulated three-patch model was needed for tracking the legs. Many problems, such as large motions, occlusions, and shadows, make motion estimation in that situation more challenging than for the facial case.

The approach we took in [7] for recognizing whole body movements was an attempt to generalize the face patch tracking technique. The basic strategy was to use the overall shape of the motion to hypothesize movements, which, in turn, proposed patch models to be verified. Our experience was that the inability to recover robustly canonical patch motion parameters made the technique brittle.

Optical flow, rather than patches, was used by Essa and Pentland [12] to estimate muscle activation on a detailed, physically-based model of the face. One recognition approach classifies expressions by a similarity measure to the typical patterns of muscle activation. Another recognition method matches motion energy templates derived from the muscle activations. These templates compress the activity sequence into a single entity. In this paper, we develop similar templates, but our templates incorporate the temporal motion characteristics.

3 Temporal Templates

Our goal is to construct a view-specific representation of movement, where movement is defined as motion over time. For now, we assume that either the background is

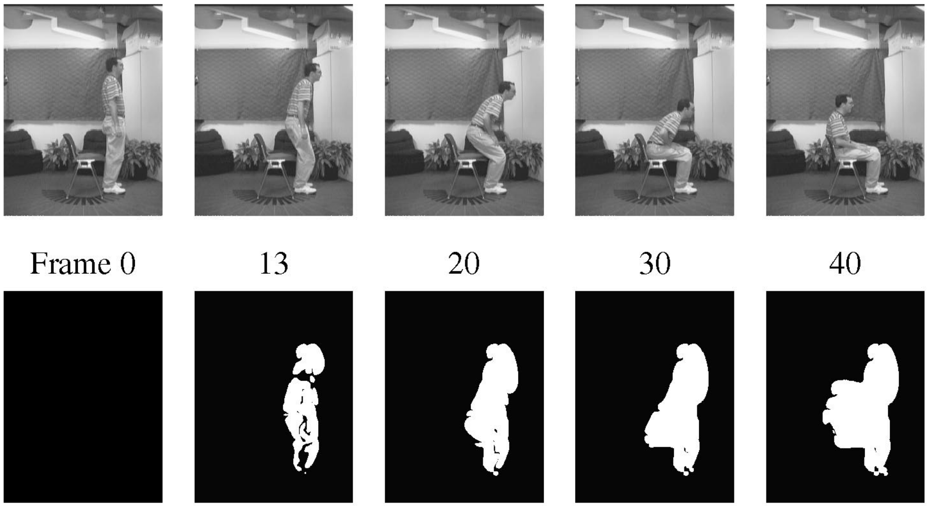


Fig. 2. Example of someone sitting. Top row contains key frames. The bottom row is cumulative motion images starting from Frame 0.

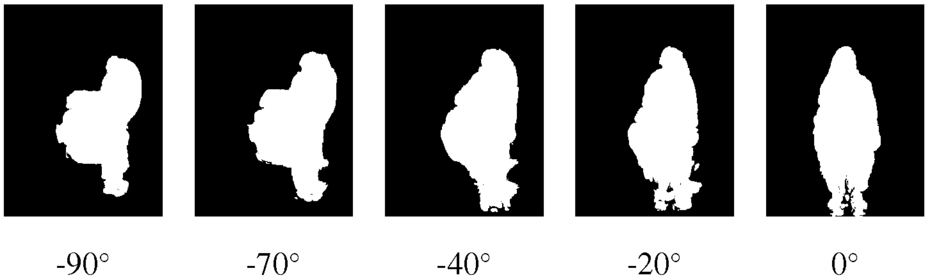


Fig. 3. MEIs of sitting movement over 90° viewing angle. The smooth change implies only a coarse sampling of viewing direction is necessary to recognize the movement from all angles.

static, or that the motion of the object can be separated from either camera-induced or distractor motion. At the conclusion of this paper, we discuss methods for eliminating incidental motion from the processing.

In this section, we define a multicomponent image representation of movement based upon the observed motion. The basic idea is to construct a vector-image that can be matched against stored representations of known movements; this image is used as a temporal template.

3.1 Motion-Energy Images

Consider the example of someone sitting, as shown in Fig. 2. The top row contains key frames in a sitting sequence. The bottom row displays cumulative binary motion images—to be described momentarily—computed from the start frame to the corresponding frame above. As expected, the sequence sweeps out a particular region of the image; our claim is that the shape of that region (*where* there is motion) can be used to suggest both the movement occurring and the viewing condition (angle).

We refer to these binary cumulative motion images as *motion-energy images* (MEI). Let $I(x, y, t)$ be an image sequence and let $D(x, y, t)$ be a binary image sequence indicating regions of motion; for many applications image-differencing is adequate to generate D. Then, the binary MEI $E_\tau(x, y, t)$ is defined

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i).$$

We note that the duration τ is critical in defining the temporal extent of a movement. Fortunately, in the recognition section we derive a backward-looking (in time) algorithm that dynamically searches over a range of τ.

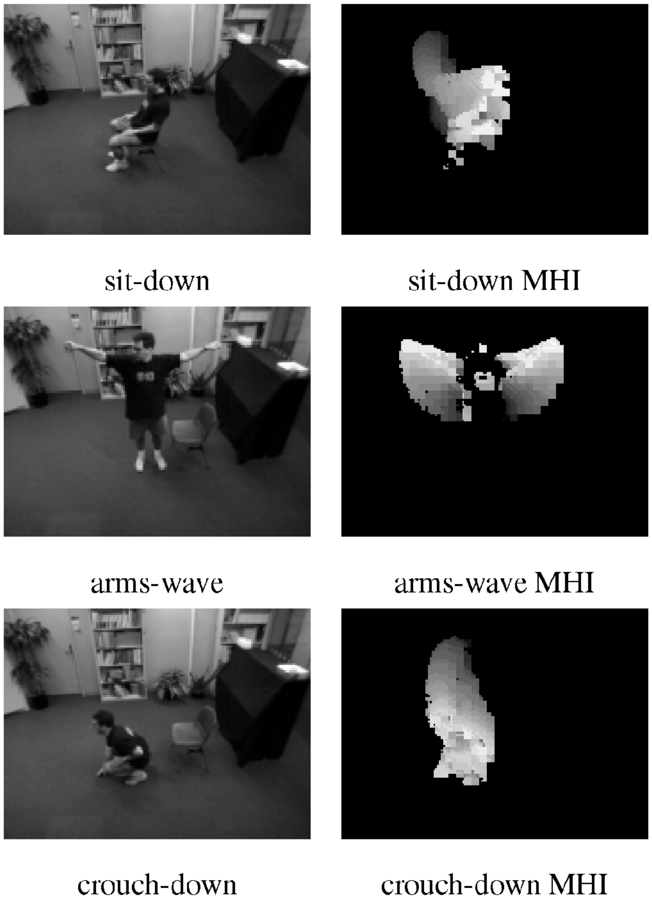
In Fig. 3, we display the MEIs of viewing a sitting motion across 90°. In [7], we exploited the smooth variation of motion over angle to compress the entire view circle into a low-order representation. Here, we simply note that because of the slow variation across angle, we only need to sample the view sphere coarsely to recognize all directions. In the evaluation section of this paper, we use 30° samplings to recognize a large variety of motions (Section 4).

3.2 Motion-History Images

To represent *how* (as opposed to where) motion the image is moving, we form a *motion-history image* (MHI). In an MHI H_τ, pixel intensity is a function of the temporal history of motion at that point. For the results presented here, we use a simple replacement and decay operator:

$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t - 1) - 1) & \text{otherwise.} \end{cases}$$

The result is a scalar-valued image where more recently moving pixels are brighter. Examples of MHIs are presented in Fig. 4.



sit-down

sit-down MHI

arms-wave

arms-wave MHI

crouch-down

crouch-down MHI

Fig. 4. Simple movements along with their MHIs used in a real-time system.

Note that the MEI can be generated by thresholding the MHI above zero. Given this situation one might consider why not use the MHI alone for recognition? To answer this question, we must wait until we describe how the MEI and MHI are used for recognizing human motion. We will show examples of how the two images together provide better discrimination than either alone.

One possible objection to the approach described here is that there is no consideration of optic flow, the direction of image motion. In response, it is important to note the relation between the construction of the MHI and direction of motion. Consider the waving example in Fig. 4 where the arms fan upwards. Because the arms are isolated components—they do not occlude other moving components—the motion-history image implicitly represents the direction of movement: The motion in the arm down position is "older" than the motion when the arms are up. For these types of articulated objects and for simple movements where there is not significant motion self-occlusion, the direction of motion is well-represented using the MHI. As motions become more complicated, the optic flow is more difficult to discern, but is typically not lost completely.

3.3 Extending Temporal Templates

The MEI and MHI are two components of a vector image designed to encode a variety of motion properties in a spatially indexed manner. Other possible components of the temporal templates include power in directional motion integrated over time (e.g., "in this pixel there has been a large amount of motion in the down direction during the integrating time window") or the spatially localized periodicity of motion (a pixel by pixel version of Polana and Nelson [22]). The vector-image template is similar in spirit to the vector-image based on orientation and edges used by Jones and Malik [18] for robust stereo matching.

For the results in this paper, we use only the two components derived above (MEI and MHI) for representation and recognition. This particular choice of temporal projection operator has the advantage that the computation is *recursive*: The MHI at time t is computed from the MHI at time $t-1$ and the current motion image $D_t(x,y)$, and the current MEI is computed by thresholding the MHI. The recursive definition implies that no history of the previous images or their motion fields need be stored nor manipulated, making the computation both fast and space efficient. Other projection operators such as pixel-wise summations over time require maintaining all the $D_t(x,y)$ for $t_0 \leq t \leq t_\tau$.

Of course, any projection operator loses information. One potential difficulty is that any interesting motion history at a given location is obliterated by recent movement. For all of our experiments, we have used the recency operator described here. We also note that such an operator has biological implications: A recency projection can be trivially performed by motion-sensitive, spatially-local filters with fixed decay rates.

4 Discrimination

4.1 Matching Temporal Templates

To construct a recognition system, we need to define a matching algorithm for the temporal template. Because we are using an appearance-based approach, we must first define the desired invariants for the matching technique. As we are using a view sensitive approach, it is desirable to have a matching technique that is as invariant as possible to the imaging situation. Therefore, we have selected a technique which is scale and translation invariant.

We first collect training examples of each movement from a variety of viewing angles. Given a set of MEIs and MHIs for each view/movement combination, we compute statistical descriptions of the these images using moment-based features. Our current choice is 7 Hu moments [17] which are known to yield reasonable shape discrimination in a translation- and scale-invariant manner (See Appendix).[1] For each view of each movement, a statistical model of the moments (mean and covariance matrix) is generated for both the MEI and MHI. To recognize an input movement, a Mahalanobis distance is calculated between the moment description of the input and each of the known movements. In this section, we show results using this distance metric in terms of its separation of different movements.

Note that we have no fundamental reason for selecting this method of scale- and translation-invariant template matching. The approach outlined has the advantage of not being computationally taxing making real-time implementation feasible. One disadvantage is that the Hu moments

[1] If required, rotation invariance (in the image plane) can be obtained as well, see the Appendix.

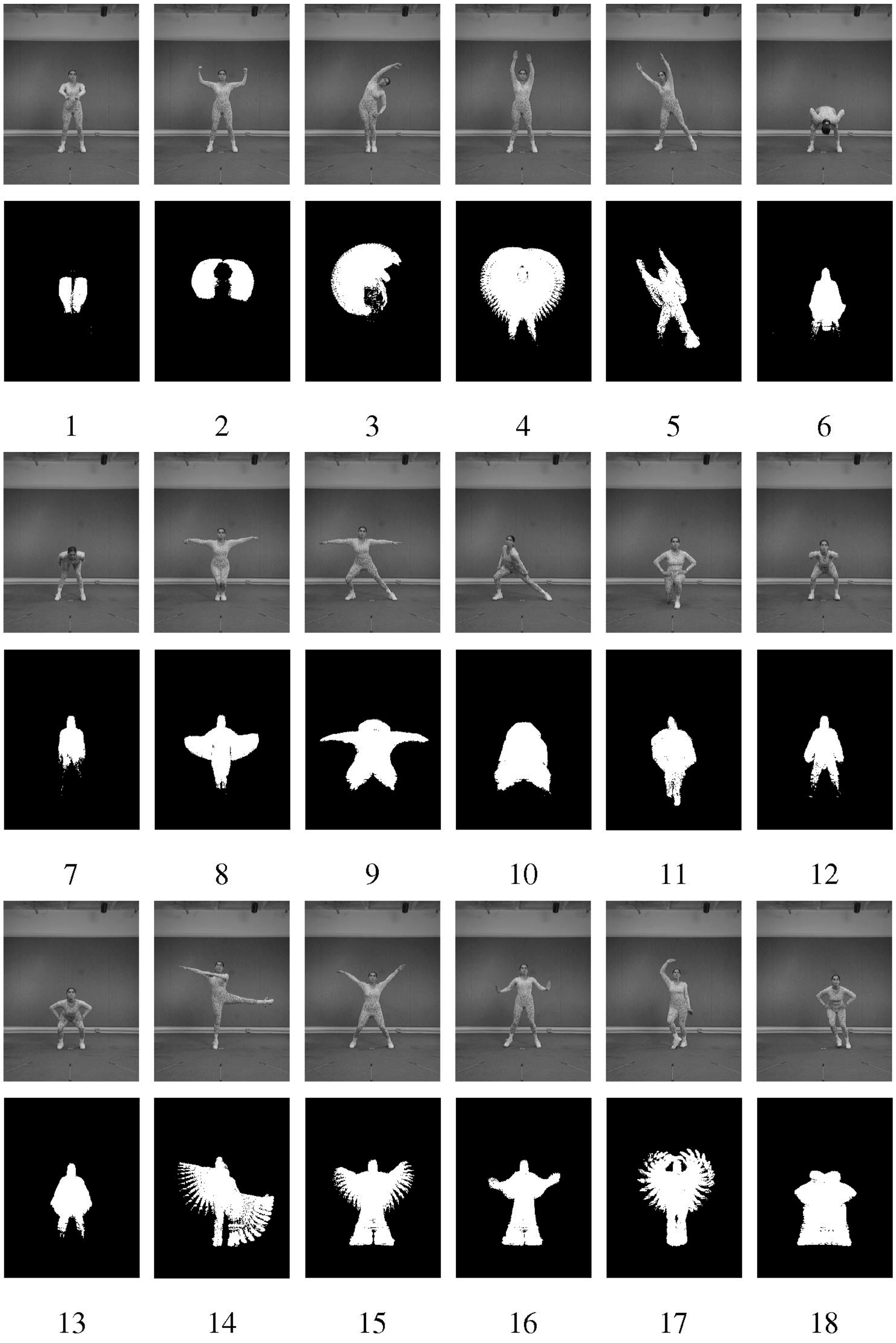


Fig. 5. A single key frame and MEI from the frontal view of each of the 18 aerobic exercises used to test the representation.

are difficult to reason about intuitively. Also, we note that the matching methods for the MEI and MHI need not be the same, in fact, given the distinction we make between where there is motion and how the motion is moving, one might expect different matching criteria.

4.2 Testing on Aerobics Data: One Camera

To evaluate the power of the temporal template representation, we recorded video sequences of 18 aerobic exercises performed several times by an experienced aerobics instructor. Seven views of the movement—$+90^\circ$ to -90° in 30° increments in the horizontal plane—were recorded. Fig. 5 shows the frontal view of one key frame for each of the moves along with the frontal MEI. We take the fact that the MEI makes clear to a human observer the nature of the motion as anecdotal evidence of the strength of this component of the representation. We also mention that the feather-like patterns seen during rapid body motion, such as the arm swing in move 14 and 17 of Fig. 5, vary in their exact phase from one instance to the next depending upon the phase relation to the image sampling. However, because the Hu moments create a coarse global shape

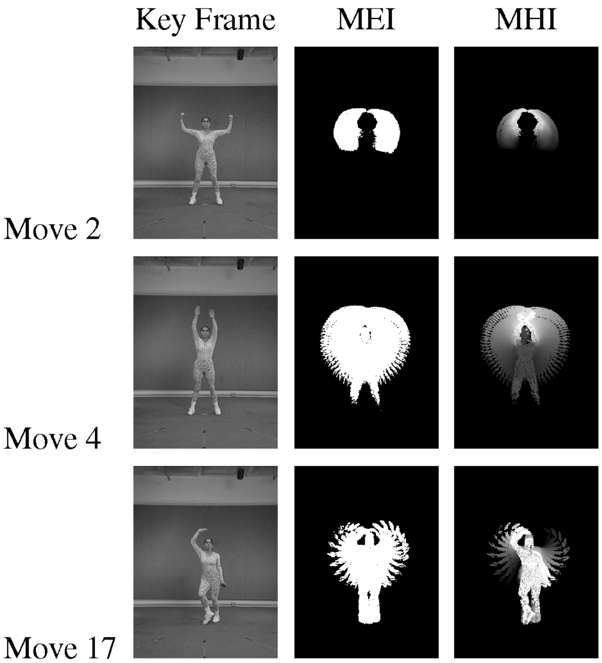


Fig. 6. Comparison of MEI and MHI. Under an MEI description moves 4 and 17 are easily confused; under the MHI, moves 2 and 4 are similar. Because the global shape descriptions are weighted by the pixel values, having both images yields more discrimination power.

TABLE 1
Test Results Using One Camera at 30° Off Frontal

	Closest	Closest	Correct	Median	Rank
	Dist	Move	Dist	Dist	
Test 1	1.43	4	1.44	2.55	2
2	3.14	2	3.14	12.00	1
3	3.08	3	3.08	8.39	1
4	0.47	4	0.47	2.11	1
5	6.84	5	6.84	19.24	1
6	0.32	10	0.61	0.64	7
Test 7	0.97	7	0.97	2.03	1
8	20.47	8	20.47	35.89	1
9	1.05	8	1.77	2.37	4
10	0.14	10	0.14	0.72	1
11	0.24	11	0.24	1.01	1
12	0.79	12	0.79	4.42	1
Test 13	0.13	6	0.25	0.51	3
14	4.01	14	4.01	7.98	1
15	0.34	15	0.34	1.84	1
16	1.03	15	1.04	1.59	2
17	0.65	17	0.65	2.18	1
18	0.48	10	0.51	0.94	4

Each row corresponds to one test move and gives the distance to the nearest move (and its index), the distance to the correct matching move, the median distance, and the ranking of the correct move.

description, the precise phase has little effect on the shape description.

The aerobics imagery provides examples of why having both the MEI and MHI is valuable even though the MEI can be constructed by thresholding the MHI. Fig. 6 shows the MEI and MHI for moves 2, 4, and 17. Note that move 4 and 17 have quite similar MEIs yet distinct MHIs; moves 2 and 4 have similar MHIs in terms of where the majority of image energy is located yet display quite distinct MEIs. Because the MEI and MHI projection functions capture two distinct characteristics of the motion—"where" and "how," respectively—the shape descriptors of these two images discriminate differently.

For this experiment, the temporal segmentation and selection of the time window over which to integrate were performed manually. Later, we will detail a self-segmenting, time-scaling recognition system (Section 5). The only preprocessing done on the data was to reduce the image resolution to 320 x 240 from the captured 640 x 480. This step had the effect of not only reducing the data set size but also of providing some limited blurring which enhances the stability of the global statistics.

We constructed the temporal template for each view of each move and then computed the Hu moments on each component. To do a useful Mahalanobis procedure would require watching several different people performing the same movements, this multisubject approach is taken in the next section where we develop a recognition procedure using a full covariance (Section 5). Instead, here we design the experiment to be a measurement of confusion. A new test subject performed each move and the input data was recorded by two cameras viewing the movement at approximately 30° to left and 60° to the right of the subject. The temporal template for each of the two views of the test input movements was constructed and the associated moments computed.

Our first test uses only the left (30°) camera as input and matches against all seven views of all 18 moves (126 total). In this experiment, we used only one instance of each view/move pair. We select as a metric a *pooled* independent Mahalanobis distance—using the same diagonal covariance matrix for all the classes as generated by pooling all the data—to accommodate variations in magnitude of the moments. Table 1 displays the results. Indicated are the distance to the move closest to the input (as well as its index), the distance to the correct matching move, the median distance (to give a sense of scale), and the ranking of the correct move in terms of least distance.

The first result to note is that 12 of 18 moves are correctly identified using the single view. This performance is quite good considering the compactness of the representation (a total of 14 moments from two correlated motion images) and the size and similarity of the target set. Second, in the typical situation in which the best match is not the correct move, the difference in distances from the input to the closest move versus the correct move is small compared to the median distance. Examples of this include test moves 1, 9, 13, 16, and 18. In fact, for moves 1, 16, and 18 the difference is negligible.

To analyze the confusion difficulties further, consider the example shown in Fig. 7. Displayed here, left to right, are the input MHI (known to be move 13 at view angle 30°), the closest match MHI (move 6 at view angle 0°), and the "correct" matching MHI of move 13. The problem is that an alternative view of a different movement projects into a temporal

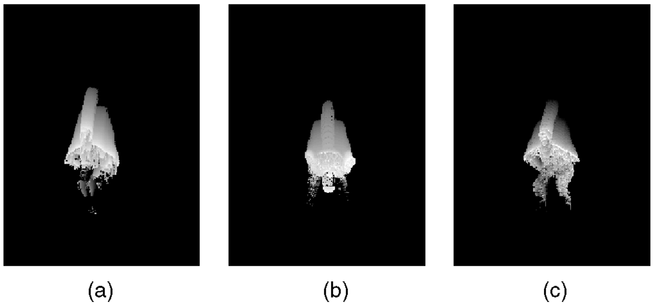


Fig. 7. An example of MHIs with similar statistics. (a) Test input of move 13 at 30°. (b) Closest match which is move 6 at 0°. (c) Correct match.

template with similar statistics. For example, consider sitting and crouching motions when viewed from the front. The observed motions are almost identical and the coarse temporal template statistics do not distinguish them well.

4.3 Combining Multiple Views

A simple mechanism to increase the power of the method is to use more than one camera. Several approaches are possible. For this experiment, we use two cameras placed such that they have orthogonal views of the subject. The recognition system now finds the minimum sum of Mahalanobis distances between the two input templates and two stored views of a movement that have the correct angular difference between them, in this case 90°. The assumption embodied in this approach is that we know the approximate angular relationship between the cameras.

TABLE 2
Results Using Two Cameras Where the Angular Interval Is Known and Any Matching Views Must Have the Same Angular Distance

	Closest Dist	Closest Move	Correct Dist	Median Dist	Rank
Test 1	2.13	1	2.13	6.51	1
2	12.92	2	12.92	19.58	1
3	7.17	3	7.17	18.92	1
4	1.07	4	1.07	7.91	1
5	16.42	5	16.42	32.73	1
6	0.88	6	0.88	3.25	1
Test 7	3.02	7	3.02	7.81	1
8	36.76	8	36.76	49.89	1
9	5.10	8	6.74	8.93	3
10	0.68	10	0.68	3.19	1
11	1.20	11	1.20	3.68	1
12	2.77	12	2.77	15.12	1
Test 13	0.57	13	0.57	2.17	1
14	6.07	14	6.07	16.86	1
15	2.28	15	2.28	8.69	1
16	1.86	15	2.35	6.72	2
17	2.67	8	3.24	7.10	3
18	1.18	18	1.18	4.39	1

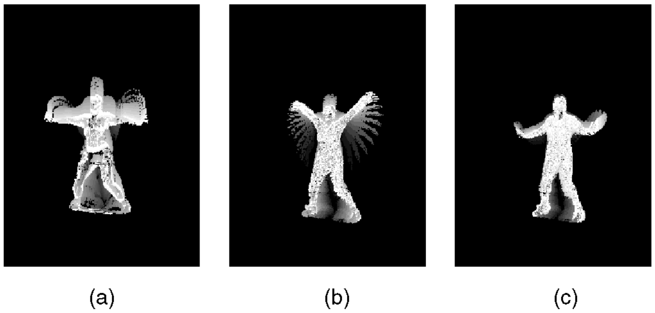


Fig. 8. Example of error where failure is caused, by both the inadequacy of using image differencing to estimate image motion and the lack of the variance data in the recognition procedure. (a) Test input of move 16. (b) Closest match which is move 15. (c) Correct match.

Table 2 provides the same statistics as the first table, but now using two cameras. Notice that the classification now contains only three errors. The improvement of the result reflects the fact that for most pairs of this suite of movements, there is some view in which they look distinct. Because we have 90° between the two input views, the system can usually correctly identify most movements.

We mention that if the approximate calibration between cameras is not known (and is not to be estimated) one can still logically combine the information by requiring consistency in labeling. That is, we remove the interangle constraint, but do require that both views select the same movement. The algorithm would be to select the move whose Mahalanobis sum is least, regardless of the angle between the target views. If available, angular order information—e.g., camera 1 is to the left of camera 2—can be included. When this approach is applied to the aerobics data shown here, we still get similar discrimination. This is not surprising because the input views are so distinct.

To analyze the remaining errors, consider Fig. 8, which shows the input for move 16. Left to right are the 30° MHIs for the input, the best match (move 15), and the correct match. The test subject performed the move much less precisely than the original aerobics instructor. Because we were not using a Mahalanobis variance across subjects, the current experiment could not accommodate such variation. In addition, the test subject moved her body slowly while wearing low frequency clothing resulting in an MHI that has large gaps in the body region. We attribute this type of failure to our simple (i.e., naive) motion analysis; a more robust motion detection mechanism would reduce the number of such situations.

5 Segmentation and Recognition

The final element of performing recognition is the temporal segmentation and matching. During the training phase, we measure the minimum and maximum duration that a movement may take, τ_{min} and τ_{max}. If the test motions are performed at varying speeds, we need to choose the right τ for the computation of the MEI and the MHI. Our current system uses a backward looking variable time window. Because of the simple nature of the replacement operator, we can construct a highly efficient algorithm for approximating a search over a wide range of τ.

The algorithm is as follows: At each time step, a new MHI $H_\tau(x, y, t)$ is computed setting $\tau = \tau_{max}$, where τ_{max} is the longest time window we want the system to consider. We choose $\Delta\tau$ to be $(\tau_{max} - \tau_{min})/(n-1)$, where n is the number of temporal integration windows to be considered.[2] A simple thresholding of MHI values less than $(\tau - \Delta\tau)$ generates $H_{(\tau-\Delta\tau)}$ from H_τ:

$$H_{\tau-\Delta\tau}(x, y, t) = \begin{cases} (H_\tau(x, y, t) - \Delta\tau) & \text{if } H_\tau(x, y, t) > \Delta\tau \\ 0 & \text{otherwise.} \end{cases}$$

To compute the shape moments, we scale H_τ by $1/\tau$. This scale factor causes all the MHIs to range from 0 to 1 and provides invariance with respect to the speed of the movement. Iterating, we compute all n MHIs; thresholding of the MHIs yields the corresponding MEIs.

After computing the various scaled MHIs and MEIs, we compute the Hu moments for each image. We then check the Mahalanobis distance of the MEI parameters against the known view/movement pairs. The mean and the covariance matrix for each view/movement pair is derived from multiple subjects performing the same move. Any movement found to be within a threshold distance of the input is tested for agreement of the MHI. If more than one movement is matched, we select the movement with the smallest distance.

The aerobics data were generated from only two individuals who performed the movements precisely, without adequate variation to generate a statistical distribution. To test the real-time recognition system, we created a new, smaller movement set using multiple people to provide training examples. Our experimental system recognizes 180° views of the movements *sitting*, *arm waving*, and *crouching* (See Fig. 4). The training required four people and sampling the view circle every 45°. The system performs well, rarely misclassifying the movements. The errors which do arise are mainly caused by problems with image differencing and also due to our approximation of the temporal search window $n < (\tau_{max} - \tau_{min} + 1)$.

The system runs at approximately 9 Hz using 2 CCD cameras connected to a Silicon Graphics 200MHz Indy; the images are digitized at a size of 160 x 120. For these three moves $\tau_{max=19}$ (approximately 2 seconds), $\tau_{min} = 11$ (approximately 1 second), and we chose $n = 6$. The comparison operation is virtually no cost in terms of computational load, so adding more movements does not affect the speed of the algorithm, only the accuracy of the recognition.

6 Extensions, Problems, and Applications

We have presented a novel representation and recognition technique for identifying movements. The approach is based upon temporal templates and their dynamic matching in time. Initial experiments in both measuring the sensitivity of the representation and in constructing real-time recognition systems have shown the effectiveness of the method.

There are, of course, some difficulties in the current approach. Several of these are easily rectified. As mentioned, a more sophisticated motion detection algorithm would increase robustness. Also, as developed, the method assumes all motion present in the image should be incorporated into the temporal templates. Clearly, this approach would fail when two people are in the field of view. To implement our real-time system, we use a tracking bounding box which attempts to isolate the relevant motions.

A worse condition is when one person partially occludes another, making separation difficult, if not impossible. Here, multiple cameras is an obvious solution. Since occlusion is view angle specific, multiple cameras reduce the chance the occlusion is present in all views. For monitoring situations, one can use an overhead camera to select which ground based cameras have a clear view of a subject and to specify (assuming loose calibration) where the subject would appear in each image.

6.1 Handling Incidental Motion

A more serious difficulty arises when the motion of part of the body is not specified during a movement. Consider, for example, throwing a ball. Whether the legs move is not determined by the movement itself, inducing huge variability in the statistical description of the temporal templates. To extend this paradigm to such movements requires some mechanism to automatically either mask away regions of this type of motion or to always include them.

For some real-time applications, such as the one discussed in the next section, we have a modified MHI generation algorithm: Instead of using the motion image to create the MHI or MEI, we simply use the background subtracted images. This way a small motion of a body part versus no motion does not change how it contributes to the temporal template. This most likely reduces the discrimination power of the system (we have not investigated this), but it does tremendously increase the robustness.

Two other examples of motion that must be removed are camera motion and locomotion (if we assume the person is performing some movement while locomoting and what we want to see is the underlying motion). In both instances, the problem can be overcome by using a body centered motion field. The basic idea would be to subtract out any image motion induced by camera movement or locomotion. Of these two phenomena, camera motion elimination is significantly easier because of the over constrained nature of estimating egomotion. Our only insight at this point is that because the temporal template technique does not require accurate flow fields it may be necessary only to approximately compensate for these effects and then to threshold the image motion more severely than we have done to date.

6.2 The KidsRoom: An Application

We conclude by mentioning an application we developed in which we employed a version of the temporal template technique described. The application was titled The KidsRoom, an interactive play-space for children [5]. The basic idea is that the room is aware of the children (maximum of 4) and takes them through a story where the responses of the room are affected by what the children do. Computers control the lighting, sound effects, performance of the score, and

[2] Ideally, $n = \tau_{max} - \tau_{min} + 1$ resulting in a complete search of the time window between τ_{max} and τ_{min}. Only computational limitations argue for a smaller n.

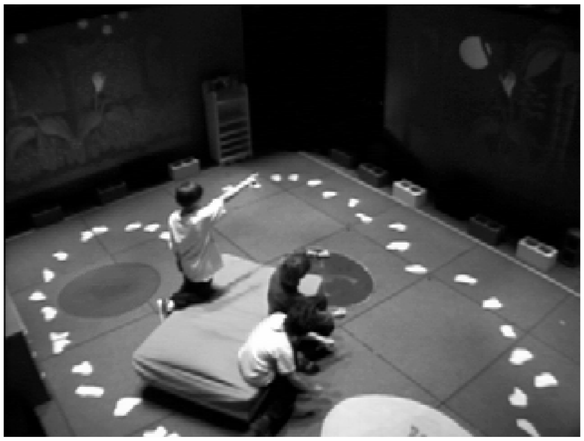


Fig. 9. The KIDSROOM interactive play-space. Using a modified version of temporal templates, the room responds to the movements of the children. All sensing is performed using vision from three cameras: one overhead for tracking and two wall mounted cameras for movement recognition.

illustrations projected on the two walls of the room that are actually video screens. The current scenario is an adventurous trip to Monsterland. A snapshot is shown in Fig. 9.

In the last scene the monsters appear and teach the children to dance—basically to perform certain movements. Using the background-subtracted modified version of the MEIs and MHIs, the room can compliment the children on well-performed moves (e.g., spinning) and then turn control of the situation over to them: the monsters follow the children if the children perform the moves they were taught. The interactive narration coerces the children to room locations where occlusion is not a problem. Of all the vision processes required, the modified temporal template is one of the more robust. We take the ease of use of the method to be an indication of its potential.

APPENDIX

IMAGE MOMENTS

The two-dimensional $(p+q)\mathrm{th}$ order moments of a density distribution function $\rho(x,y)$ (e.g., image intensity) are defined in terms of Riemann integrals as:

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q \rho(x,y)\,dx\,dy, \tag{1}$$

for $p, q = 0, 1, 2, \cdots$.

The central moments μ_{pq} are defined as:

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q \rho(x,y)\,d(x-\bar{x})\,d(y-\bar{y}), \tag{2}$$

where

$$\bar{x} = m_{10}/m_{00},$$
$$\bar{y} = m_{01}/m_{00}.$$

It is well-known that under the translation of coordinates, the central moments do not change, and are therefore invariants under translation. It is quite easy to express the central moments μ_{pq} in terms of the ordinary moments m_{pq}. For the first four orders, we have

$$\begin{aligned}
\mu_{00} &= m_{00} \equiv \mu \\
\mu_{10} &= 0 \\
\mu_{01} &= 0 \\
\mu_{20} &= m_{20} - \mu\bar{x}^2 \\
\mu_{11} &= m_{11} - \mu\bar{x}\bar{y} \\
\mu_{02} &= m_{02} - \mu\bar{y}^2 \\
\mu_{30} &= m_{30} - 3m_{20}\bar{x} + 2\mu\bar{x}^3 \\
\mu_{21} &= m_{21} - m_{20}\bar{y} - 2m_{11}\bar{x} + 2\mu\bar{x}^2\bar{y} \\
\mu_{12} &= m_{12} - m_{02}\bar{x} - 2m_{11}\bar{y} + 2\mu\bar{x}\bar{y}^2 \\
\mu_{03} &= m_{03} - 3m_{02}\bar{y} + 2\mu\bar{y}^3.
\end{aligned}$$

To achieve invariance with respect to orientation and scale, we first normalize for scale defining η_{pq}:

$$\eta_{pq} = \frac{\mu_{pq}}{(\mu_{00})^{\gamma}},$$

where $\gamma = (p+q)/2 + 1$ and $p + q \geq 2$. The first seven orientation invariant Hu moments are defined as:

$$\begin{aligned}
\nu_1 &= \eta_{20} + \eta_{02} \\
\nu_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\nu_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\nu_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\nu_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\
&\quad \cdot [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\nu_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\nu_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2].
\end{aligned}$$

These moments can be used for pattern identification independent of position, size, and orientation.

ACKNOWLEDGMENTS

Portions of this paper appeared at the IEEE Conference on Computer Vision and Pattern Recognition 1997, San Juan [11]. This research was conducted while both authors were at the MIT Media Laboratory and was funded in part by the Digital Life consortium. We also acknowledge the anonymous reviewers for comments that lead to clarification of the paper.

REFERENCES

[1] J. Aggarwal and N. Nandhakumar, "On the Computation of Motion of Sequences of Images—A Review," *Proc. IEEE*, vol. 69, no. 5, pp. 917-934, 1988.

[2] J. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428-440, 1999.

[3] K. Akita, "Image Sequence Analysis of Real World Human Motion," *Pattern Recognition*, vol. 17, no. 1, pp. 73-83, 1984.

[4] M.J. Black and Y. Yacoob, "Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using Local Parametric Model of Image Motion," *Proc. Int'l Conf. Computer Vision*, pp. 374-381, 1995.

[5] A.F. Bobick, S.S. Intille, J.W. Davis, F. Baird, L.W. Campbell, Y. Ivanov, C.S. Pinhanez, A. Schütte, and A. Wilson, "The Kids-Room: A Perceptually-Based Interactive and Immersive Story Environment," *Presence*, vol. 8, no. 4, pp. 368-393, Aug. 1999.

[6] A. Bobick, "Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion," *Philosophical Trans. Royal Soc. London,* vol. 352, pp. 1257-1265, 1997.

[7] A. Bobick and J. Davis, "An Appearance-Based Representation of Action," *Proc. Int'l Conf. Pattern Recognition,* pp. 307-312, 1996.

[8] L. Campbell and A. Bobick, "Recognition of Human Body Motion Using Phase Space Constraints," *Proc. Int'l Conf. Computer Vision,* pp. 624-630, 1995.

[9] Y. Cui, D. Swets, and J. Weng, "Learning-Based Hand Sign Recognition Using Shoslif-m," *Proc. Int'l Conf. Computer Vision,* pp. 631-636, 1995.

[10] T. Darrell and A. Pentland, "Space-Time Gestures," *Proc. Computer Vision and Pattern Recognition,* pp. 335-340, 1993.

[11] J. Davis and A. Bobick, "The Representation and Recognition of Human Movement Using Temporal Templates," *Proc. Computer Vision and Pattern Recognition,* pp. 928-934, 1997.

[12] I. Essa and A. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 757-763, July 1997.

[13] W. Freeman and M. Roth, "Orientation Histogram for Hand Gesture Recognition," *Proc. Int'l Workshop Automatic Face and Gesture Recognition,* pp. 296-301, 1995.

[14] D.M. Gavrila and L.S. Davis, "3D Model-Based Tracking of Humans in Action: A Multiview Approach," *Proc. Computer Vision and Pattern Recognition,* pp. 73-80, 1996.

[15] L. Goncalves, E. DiBernardo, E. Ursella, and P. Perona, "Monocular Tracking of the Human Arm in 3D," *Proc. Int'l Conf. Computer Vision,* pp. 764-770, Aug. 1995.

[16] D. Hogg, "Model-Based Vision: A Paradigm to See a Walking Person," *Image and Vision Computing,* vol. 1, no. 1, pp. 5-20, 1983.

[17] M. Hu, "Visual Pattern Recognition by Moment Invariants," *IRE Trans. Information Theory,* vol. 8, no. 2, pp. 179-187, 1962.

[18] D. Jones and J. Malik, "Computational Framework for Determining Stereo Correspondence from a Set of Linear Spatial Filters," *Image and Vision Computing,* vol. 10, no. 10, pp. 699-708, 1992.

[19] S. Ju, M. Black, and Y. Yacoob, "Cardboard People: A Parameterized Model of Articulated image Motion," *Proc. Second Int'l Conf. Automatic Face and Gesture Recognition,* pp. 38-44, Oct. 1996.

[20] J. Little and J. Boyd, "Describing Motion for Recognition," *Int'l Symp. Computer Vision,* pp. 235-240, Nov. 1995.

[21] P. Maes, T. Darrell, B. Blumberg, and A. Pentland, "The ALIVE System: Wireless, Full-Body Interaction with Autonomous Agents," *ACM Multimedia Systems,* 1996.

[22] R. Polana and R. Nelson, "Low Level Recognition of Human Motion," *Proc. IEEE Workshop Non-Rigid and Articulated Motion,* pp. 77-82, 1994.

[23] J. Rehg and T. Kanade, "Model-Based Tracking of Self-Occluding Articulated Objects," *Proc. Int'l Conf. Computer Vision,* pp. 612-617, 1995.

[24] K. Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences," *CVGIP, Image Understanding,* vol. 59, no. 1, pp. 94-115, 1994.

[25] E. Shavit and A. Jepson, "Motion Understanding Using Phase Portraits," *Proc. IJCAI Workshop: Looking at People,* 1993.

[26] J.M. Siskind, "Grounding Language in Perception," *Artificial Intelligence Rev.,* vol. 8, pp. 371-391, 1995.

[27] A. Wilson and A. Bobick, "Learning Visual Behavior for Gesture Analysis," *Proc. IEEE Int'l. Symp. Computer Vision,* Nov. 1995.

[28] Y. Yacoob and L. Davis, "Recognizing Human Facial Expressions Form Long Image Sequences Using Optical Flow," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, pp. 636-642, 1996.

[29] J. Yamato, J. Ohya, and K. Ishii, "Recognizing Human Action in Time Sequential Images Using Hidden Markov Models," *Proc. Computer Vision and Pattern Recognition,* pp. 379-385, 1992.

Aaron F. Bobick received the PhD degree in cognitive science from the Massachusetts Institute of Technology in 1987 and has also received the BS degrees from MIT in mathematics and computer science. In 1987, he joined the Perception Group of the Artificial Intelligence Laboratory at SRI International and soon after was jointly named a visiting scholar at Stanford University. From 1992 until July 1999, he served as an assistant and, then, associate professor in the Vision and Modeling Group of the MIT Media Laboratory. In 1999, Dr. Bobick moved to the College of Computing at the Georgia Institute of Technology, where he is an associate professor and serves as acting director of the GVU Center. He has performed research in many areas of computer vision. His primary work has focused on video sequences where the goal is to understand the activity in the scene. He has published papers addressing many levels of the problem from validating low level optic flow algorithms to constructing multirepresentational systems for an autonomous vehicle to the representation and recognition of human activities. The current emphasis of his work is understanding human behavior in context, especially the context of the "Aware Home" project at the Georgia Tech Broadband Institute Residential Laboratory. He is a member of the IEEE Computer Society.

James W. Davis received the MS (1996) and PhD (2000) degrees from Massachusetts Institute of Technology Media Laboratory specializing in computational vision. Currently, he is an assistant professor in the Computer and Information Science Department at The Ohio State University, Columbus. His research interests include the representation and recognition of human motion, categorization of animate motions, gesture recognition, and human-computer interactive systems. He is a member of the IEEE Computer Society.